

Методология создания виртуального Музея науки и техники СО РАН

А.М. Федотов, Ю.И. Молородов
Институт вычислительных технологий СО РАН
e-mail: fedotov@ict.nsc.ru, yumo@ict.nsc.ru

1 Введение

В последнее время в Internet-сообществе образовалась и стала популярной идея создания виртуальной организации научного взаимодействия в сфере использования информационно-вычислительных ресурсов и коммуникаций (GRID-технология). Одна из главных задач движения GRID состоит в предоставлении сетевому сообществу новых средств взаимодействия, формирование виртуальных альянсов специалистов с целью получения реальных результатов. Основное направление – “распределенное управление ресурсами” – предназначено для консолидации распределенных информационно-вычислительных ресурсов.

Необходимо разработать технологии, основанные на идеологии GRID для создания распределенных информационных систем, предназначенных для хранения и отображения информации различного уровня, а так же создать специализированные распределенные информационные системы для представления электронных коллекций научно-технической тематики. При этом информация может быть представлена самыми различными ресурсами: базы (таблицы) данных, идеографическая информация (описания), аудио- и видеофрагменты, изображения (фотографии и пр.), результаты наблюдений и списки данных, модели и модельные данные. Ресурсы могут принадлежать разным институтам, каждый из которых проводит самостоятельную политику в отношении их использования доступа. Одновременно все эти ресурсы объединяются в единую библиотеку, что открывает ее пользователям всю информацию, накопленную за долгое время, позволяет быстро получать исчерпывающие ответы на сложные запросы.

При этом фрагменты данных, выбранные по запросу, подвергаются определенным процедурам, выполняемым специализированным программным обеспечением, что позволяет представить их пользователю в адекватной форме. Решение задачи предусматривает не просто обмен документами, но и прямой одновременный доступ к программам, данным и иным ресурсам. Неоднородность хранения информации по разным научным центрам СО РАН качественно отличаются от известных моделей распределенного взаимодействия и требуют создания новых инструментов. Так, данные составляют ценность лишь тогда, когда они соответствующим образом организованы и доступны для всех участников проекта.

2 Постановка задачи

Для создания полнофункциональной информационной системы необходимо решить следующие задачи:

Сформулировать концепцию отбора информационных документов (музейных экспонатов и материалов), а также принципы формирования виртуального музейного фонда.

Определить набор интерфейсов доступа к документам музея.

Определить технологические стандарты интеграции информационных ресурсов, представляемых различными участниками проекта, и интеграции с уже существующими системами.

Сформулировать требования к лингвистическому процессору системы, определить порядок формирования онтологии музея, словарей, тезаурусов, классификаторов, каталогов и других информационных структур.

Сформировать базовые информационные структуры (БИС) для представления документов и музейных материалов (экспонатов).

Разработать архитектуру распределенной информационной системы, способной устойчиво и непрерывно функционировать в глобальной сети.

Разработать компоненты программного обеспечения для поддержки функционирования распределенной системы.

Провести наполнение баз данных документами (музейными материалами).

Но, прежде всего, необходимо разработать технологии создания распределенных информационных систем, предназначенных

для хранения и отображения разнородной информации, предназначенной для системы представления электронных коллекций научно-технической тематики. Определить стандарты передачи информации для интеграции существующих ресурсов. Выбрать основные содержательные компоненты, составляющие информационное наполнение системы интеграция их в уже функционирующие информационные системы СО РАН. Но, главное – провести инвентаризацию основных ресурсов, которые могут составить информационное наполнение системы.

Принципы технологии.

1. Необходимо отметить, что электронные документы (экспонаты) Музея носят разноплановый характер: это базы (таблицы) данных и идеографическая информация (описания, книги и статьи), изображения (фотографии и пр.) и видео- (аудио-) информация, библиографическая информация, списки данных, модели и модельные данные и др. Эти ресурсы, как правило, принадлежат различным институтам и организациям, которые проводят самостоятельную политику в отношении их использования и публичного доступа к ним. Организация интегрированной системы доступа к данным требует, помимо решения задачи модификации первичных ресурсов организаций в направлении их частичного открытия внешним пользователям, разработки принципов организации ресурсов внутреннего пользования и поддержки метаданных.

Виртуальный музей как распределенная информационная система должен обладать рядом свойств, обеспечивающих ее устойчивое функционирование. В этой связи были сформулированы следующие технологические требования к системе, а именно:

Узлы системы должны обеспечивать автономное функционирование. Выход из строя некоторых узлов системы не должен влиять на работу узла, на котором нет отказов. Все узлы должны рассматриваться как равные. Например, пользователи узла в Томске при обращении к музейным материалам из Томска не должны заметить “ремонтных работ” на узлах Новосибирска или Иркутска.

Система должна обладать свойством “прозрачности расположения”. Пользователь не должен знать о физическом месте размещения необходимой ему информации.

Архитектура системы должна обеспечивать независимость системы от типов используемых СУБД, операционных систем, аппаратных платформ и сетевых протоколов

2. Была разработана технология создания интегрированных распределенных информационных систем (ИРИС). Технология базируется на концепциях информационной модели документа, унифицированных описаний и интерфейсов доступа к информации на основе метаинформации и трехуровневой модели представления данных и метаданных. Под понятием “Документ” здесь понимается следующее. В информационном пространстве события, факты и любые другие сущности реального или виртуального мира существуют только в форме Документов. Вследствие этого Документ является основным “объектом”, с которым оперирует любая информационная система. Таким образом, Документ это любое описание реальной сущности (объекта, факта или понятия, включая алгоритмы и программы), которые составляют информационное наполнение системы.

Создание распределенной информационно-вычислительной сети опирается на идею электронных (цифровых) библиотек. В рамках этого подхода цифровые библиотеки рассматриваются как отдельная конкретная технология работы с информацией. Основной принцип, закладываемый в технологию цифровых библиотек, заключается в унифицированной каталогизации всех типов информационных ресурсов, представленных в библиотеке, т.е. информационных ресурс становится Документом тогда и только тогда, когда к нему добавлено в соответствии со стандартами функционирования открытых систем (OSI) соответствующее метаописание. Цифровые библиотеки можно рассматривать как пример виртуальной организации — большая часть компонентов находится в разных местах, при выполнении иных задач может функционировать независимо, интероперабельность достигается использованием согласованного набора протоколов и служб. Единство достигается, прежде всего, за счет унифицированного интерфейса и централизованного администрирования.

Для разработки распределенных информационных систем сегодня используются различные технологии (RPC, DCOM, RMI, ODBC, JDBC, CORBA, SOAP, eXML, WSDL, WSFL, UDDI и др.). Большая часть этих технологий по своей сути ориентирована не на работу с информационными ресурсами, а на сетевое взаимодействие программ и распределенные вычисления в гетерогенных средах. Однако ни одна из упомянутых технологий не обеспечивает универсальных способов работы с информационными ресурсами.

ми, т.е. необходимого уровня абстрагирования от конкретных систем и платформ при сохранении высокой степени функциональности и предоставления механизма доступа к данным. Эти требования присущи технологии, в основе которой лежит протокол ANSI Z39.50 (стандарт ISO-23950).

В основе протоколов семейства OSI находится поддержка метаданных -формализованных знаниях о внутренней структуре и поведении информационных и вычислительных ресурсов (или документов): структурированные сведения о ресурсе, представляющие его свойства (атрибуты) и функции – информации, предназначенной для анализа, дизайна, развития, обработки поиска и использования Документов.

Понятие метаданных уже давно и успешно используют в таких контекстах, как информационные хранилища и системы аналитической обработки данных, электронный документооборот и управление потоками работ, управление знаниями.

В основе реализации ИРИС лежит метамодель, исходящая из того, что документ (информационный ресурс: документ, алгоритм, программа, файл или вычислительный ресурс) характеризуется набором присущих ему атрибутов и методов, характеризующих связи с другими документами. Информация о документах системы, их атрибутах и методах поддерживается сервером метаданных, содержащим метаописания системы и метаописания отдельных коллекций документов. Сервер метаданных является отдельной частью системы, содержащей описание информационной модели предметной области, параметров настройки стандартных функций системы. По информации сервера метаданных осуществляется динамическая генерация схем базы данных системы и ведение программного обеспечения служебных баз данных, в которых хранятся данные, обеспечивающие поддержку стандартных функций системы, динамически определяемые отношения между документами и динамическое распределение ресурсов системы.

ИРИС обеспечивает систему работы с документами различного происхождения (объединение распределенных и локальных электронных информационных и программно-алгоритмических ресурсов, включая документооборот), систему электронной поддержки сбора и накопления информации (системы электронных коллекций, баз данных и т.п.). Принципы, заложенные в проектирование системы, позволяют автоматизировать процессы создания элек-

тронных коллекций, библиотек и т.п.

Цифровая библиотека ИРИС опирается на онтологию, описывающую способ организации информации. Онтология конкретизируется в схеме атрибутов, описывающих информацию в конкретных метаданных. ИРИС должна предоставлять следующий набор инструментов (служб) для манипуляции с наборами данных:

публикация/регистрация новых наборов данных;

база метаданных для поиска данных по атрибутам;

доступ к гетерогенным ресурсам посредством брокера ресурсов;

контроль аутентификации и доступа;

мониторинг вычислительных ресурсов и ресурсов ввода/вывода;

распределенное исполнение служб.

Соответствующие службы регистрируются в объектно-ориентированной базе данных и вызываться для обработки любого набора данных, хранящегося в библиотеке. Комбинация возможности доступа к данным через базу метаданных, служб каталогов, зарегистрированного набора методов обработки данных позволяет решать задачи для создания среды обработки данных.

Основное назначение ИРИС связано с созданием единой распределенной информационной среды Отделения, объединяющей в интегрированное информационное пространство распределенных и локальных электронных ресурсов (информационных, программных, алгоритмических) организаций Отделения и комплекса программно-технических средств, обеспечивающего использование этих ресурсов и полнофункциональное управление ими. Создание системы связано с информационной поддержкой исследований по фундаментальным и прикладным направлениям, проводимым в институтах Отделения, а также межинститутских междисциплинарных научных исследований. Создаваемая система призвана обеспечить поддержку единой распределенной информационной среды Отделения, а именно:

Информационную поддержку исследований по фундаментальным и прикладным направлениям, проводимым в институтах Отделения, а также межинститутских междисциплинарных научных исследований.

Поддержку проблемно-ориентированных систем подготовки и обмена научных ресурсов и документов с элементами удаленной совместной работы. Единый инструмент для ведения каталогов ресурсов.

Поддержку профессионально-ориентированных систем доступа и интерфейсов с банками данных и автоматизированными библиотеками.

Поддержку перспективных систем телекооперации исследователей на базе современных технологий.

В настоящее время в системе поддерживаются следующие типы информационных ресурсов:

Каталоги информационных ресурсов, т.е. информация об информационных ресурсах, или метаинформация. К этому типу следует отнести всю вторичную информацию об информационных ресурсах, в том числе электронные каталоги традиционных библиотек, электронные версии реферативных журналов, каталоги музеев, архивов и т.п. К этому типу также следует относить описания ресурсов WEB-серверов, базы данных по организациям, сотрудникам и т.д.

Электронные коллекции, т.е. совокупности цифровых объектов, объединенных по каким-либо общим признакам. Электронные коллекции представляют собой первичный информационный ресурс. Описания этих коллекций и отдельных цифровых объектов являются ресурсами вторичными, т.е. каталогами.

Классификационные ресурсы, т.е. различные электронные рубрикаторы, тезаурусы, схемы и справочники, нормирующие правила составления каталогов информационных ресурсов и отношения между цифровыми объектами.

Хранилища программного обеспечения, которое обеспечивает функционирование ИРИС в целом, ее отдельных компонент и информационных ресурсов.

Административные информационные ресурсы, включающие полную информацию об актуальном состоянии ИРИС и ее отдельных компонент.

В основу создания ИРИС положен принцип информационных хранилищ, с учетом поддержки уже функционирующих технологий, технологические решения используют принципы обработки метаописаний данных на основе серверов метаданных и протокола Z39.50. Разработана технология, которая обеспечивает виртуальную интеграцию разнородных информационных и вычислительных ресурсов, расположенных на серверах различных организаций, в единую систему с унифицированным доступом на основе открытых международных стандартов. Это позволит объединить в

единое информационное пространство уже существующие многочисленные сервера организаций, входящих в научно-образовательную сеть, основанные на стандартных протоколах, организовать информационное обеспечение проведения исследований по фундаментальным и прикладным направлениям, проводимым в институтах, а также обеспечить доступ к фактографическим базам данных и уникальным коллекциям, накопленным в научных подразделениях СО РАН, с помощью современных средств доступа к данным.

3. Разработаны корпоративные стандарты описания метаданных, интерфейсы сопряжения разнородной информации, хранящейся в различных типах баз данных, и единый интерфейс актуализации информации.

На основе существующих стандартов разработана технология, позволяющая описывать классы объектов, представляющие типы информационных документов, атрибуты и отношения между различными типами документов и ресурсов. Разработанная технология уже сейчас позволяет создать систему унифицированного доступа к документам (прототип системы уже функционирует на WWW сервере СО РАН). Решение задачи предусматривает не просто обмен документами, но и прямой одновременный доступ к программам, данным и иным ресурсам. Доступ должен предоставляться на определенных условиях, формулируемых владельцем ресурса (система обеспечения прав доступа к ресурсам, динамической авторизации, определения доступных обобщенных ресурсов, их качества, условий использования пока находится в стадии разработки).

Разработаны основные положения по созданию корпоративных стандартов описания метаданных и интерфейсов сопряжения разнородной информации, хранящейся в различных типах баз данных. Разработанные положения обосновывают технологические решения, обеспечивающих виртуальную интеграцию описаний разнородных информационных ресурсов, расположенных на серверах различных организаций, в единую базу данных на основе открытых международных стандартов. Разработанные положения позволили создать единый интерфейс актуализации информации. Данный интерфейс уже работает в административной части ИРИС.

Система обмена метаданными основывается на сервере метаданных СО РАН, который на данный момент поддерживает следу-

ющий набор служб: публикация/регистрация новых наборов данных; база метаданных для поиска данных по атрибутам; доступ к гетерогенным ресурсам посредством брокера ресурсов.

Разработанные технологии реализуются в рамках ИРИС. Пользователю система (на первом этапе) предоставляет: каталоги организаций и сотрудников; интегрированный каталог основных разработок, проектов, публикаций и изданий, включая электронные; каталоги документов и ресурсов, касающиеся всего научно-образовательного пространства; каталоги предоставляемых программных, алгоритмических и вычислительных ресурсов; доступ к электронной библиотеке – хранилищу ресурсов.

4. Разработаны подходы и технологии, обеспечивающие виртуальную интеграцию описаний разнородных информационных ресурсов, расположенных на серверах различных организаций, в единую базу данных на основе открытых международных стандартов.

В рамках работ по проекту разработана и введена в опытную эксплуатацию система поддержки распределенных метаданных (каталогов) информационных ресурсов (МИР). Описания метаданных, основывается иерархической схеме данных OSI. В качестве головной схемы данных выбрана схема "Дублинское ядро" (Dublin Core). Схема данных - это наименование правил, в соответствии с которыми описывается содержание данного ресурса. Эти схемы включают контролируемые словари, системы записи или правила анализа. Значение элементов, определенное с помощью схемы, представляет собой символ (набор символов), выбранный из контролируемого словаря (например, индекс системы классификации или набор предметных рубрик), либо строку определенной структуры.

В настоящее время в основе большинства систем метаданных лежат расширенные спецификации Дублинского ядра (DC). Элементы DC входят практически во все системы метаданных, но в чистом виде практически не используется. Так как стандарт DC определяет минимальный набор элементов для описания ресурсов, то возникает необходимость его дополнения собственным набором элементов, в общем случае, достаточно большим. Основные элементы метаданных DC разбиваются на 3 группы, которые соответствуют классу или области информации, хранящейся в них: элементы, относящиеся к описанию содержания ресурсов (Content), элементы, относящиеся к интеллектуальной собствен-

ности (Intellectual Property) и элементы, относящиеся к идентификации ресурсов (Instantiation). Спецификация DC включает 15 основных элементов. Элементы могут повторяться и кроме этого, элемент может разбиваться на подэлементы. В настоящее время стандартизирован только набор полей DC. Ряд элементов имеет подэлементы. При описании элемента вводятся понятия схемы и подэлементов (элемент может представлять собой ссылку на другую схему данных).

Словарь системы МИР в настоящий момент содержит немногим более 40 элементов, который в случае необходимости может быть расширен (расширение будет проводится в соответствии со схемами представления метаданных в схемах GILS, CIMI и DigitalCollections.). Как показал опыт по наполнению системы, что данный набор элементов является вполне достаточным для первого этапа работы по созданию каталогов (метаописаний) электронных документов.

Технология интеграции ресурсов основана на открытом международном стандарте Z39.50 (ISO23950), определяющем сетевой доступ к базам данных. Стандарт Z39.50 включает в себя все необходимые компоненты для организации распределенной информационной системы, основанной на жестких (глобально стандартизованных) правилах каталогизации. В системе реализованы два варианта доступа к данным на “физическом” уровне: первый реализует унифицированные SQL запросы к реляционным базам данным (Type 104 Z39.50), второй реализует запросы в соответствии с протоколом ANSI Z39.50. Модернизация серверного программного обеспечения Z39.50 (ZooPARK) для работы с метаданными в схемах CIMI и DigitalCollections совместно с другими схемами.

В связи с этим была произведена модернизация шлюза Z39.50-WWW (ZGW) для поиска и представления метаданных в схемах CIMI и DigitalCollections в распределенных информационных системах на основе Z39.50. Создание пользовательских интерфейсов общего назначения. Подготовка тестовой записи в формате XML с семантическим наполнением полей CIMI, DigitalCollections и Dublin Core. Проработка технологии отображения реальных данных (система КАМИС) по культурному наследию (недвижимые памятники и живопись) в структуры CIMI и предоставление доступа к ним по Z39.50.

Наиболее важный нерешенный вопрос это создание онтологии (для каждой предметной области – математика, химия,

механика, геология и др.) и поддержка словарей-классификаторов предметной области, которые необходимы для упорядочения информации, представленной в каталогах, и создания систем поиска информации (СПИ).

5. Следующим вопросом задачи интеграции информационных ресурсов двух Отделений является вопрос о содержательном наполнении каталогов на текущем этапе.

Содержательное наполнение распределенного каталога должно вести по двум основным направлениям:

Электронные документы

Документы нашего научно-культурного наследия.

Электронные документы в первую очередь должны быть интегрированы с уже ведущимися в СО РАН информационными системами, которые включают:

Административные ресурсы:

информация о структуре Отделений, о структуре ее подразделений,

адресно-справочные сведения; персоналии ведущих ученых
нормативно-правовые документы;

Научные информационные ресурсы:

научные публикации ученых (включая существующие системы ведения научных трудов ГПНТБ, ОИГГМ СО РАН и СОЦИОНЕТ ИЭОПП СО РАН),

труды научных конференций;

научные издания Отделений;

информация о научных проектах, конкурсах и грантах;

информация о научных коллективах (включая биографии и статьи об ученых, института и разработках);

информация об основных разработках и достижениях СО РАН;

Описания электронных коллекций (включая данные спутникового мониторинга).

Описания научных коллекций.

Описания программных и алгоритмических ресурсов.

Документы научно-культурного наследия, описанные в принятых схемах данных отображают памятники научной деятельности СО РАН. Все памятники науки и техники разделяются по способу кодирования заключенной в них информации на следующие группы:

вещественные (движимые и недвижимые);

письменные;
изобразительные (включая фото-, кино-, аудио-, и видеопоматники);
памятные места.

Среди вещественных, например, необходимо выделить такие виды:

научные приборы или коллекции, имеющие фундаментальное значение и связанные с работой больших научных коллективов, лабораторий и других подразделений

приборы, установки и коллекции историко-научного значения, собранные научным коллективом под руководством крупнейших ученых

памятники, характеризующие практическое использование фундаментальных научных достижений

инструменты и приспособления, которыми пользовались в своих исследованиях известные ученые, инженеры, изобретатели и т.д.

6. Разработаны технологические и поведенческие модели информационного взаимодействия пользователей.

7. Исследованы проблемы ведения распределенных баз данных и коллекций. Разработаны система электронных публикаций данных произвольной природы, развиты способы использования метаданных для поисковых целей, разработаны технологии создания и обслуживания электронных библиотек. Система поддержки электронных коллекций (АТЛАС) эксплуатируется уже на протяжении 4 лет. В основе этой системы также лежат метаданные, описывающие структурные свойства документов, составляющих содержательное наполнение коллекций. В данной системе эксплуатируется более 50 коллекций, большая часть которых относится к вопросам описания биразнообразия и некоторых других областей наук о жизни (биологии).

3 Заключение

Для дальнейшей разработки принципов необходимо продолжить работу по следующим направлениям.

- Развитие серверов метаданных и серверов каталогов (отработка технологии использования и усовершенствование интерфейсов).

- Реализация программ и интеллектуальных агентов в составе серверов.
- Провести инвентаризацию имеющихся информационных ресурсов СО РАН, которые должны быть включены в коллекции музея (включая архив газеты “Наука в Сибири”, коллекции фотоматериалов и др.).
- Наполнение каталогов Документами и Электронными экспонатами.

Список литературы

- [1] Баракхин В.Б., Федотов А.М. Построение тезауруса для информационно-поисковой системы "Web-ресурсы математического содержания" // Инфокоммуникационные и вычислительные технологии и системы. Материалы Всероссийской конференции. Улан-Удэ — Байкал, 5–9 августа 2003 г. Улан-Удэ, издательство БурГУ. 2003. С. 21–23.
- [2] Жижимов О.Л., Мазов Н.А. Как построить распределенную систему на основе протокола Z39.50 // Материалы выездного заседания координационного научного совета СО РАН по целевой программе “Информационно-телекоммуникационные ресурсы СО РАН”. – Иркутск. – 2003. – С. 16-20.
- [3] Леонова Ю.В., Федотов А.М. О содержательном наполнении системы информационного обеспечения научного сообщества // Вычислительные технологии, Спец. выпуск, Т.8, 2003, Материалы Международной конференции “Вычислительные и информационные технологии в науке, технике и образовании”, Усть-Каменогорск, 11-14 сентября 2003 г., Ч.4, С. 294-303.52 с.
- [4] Молородов Ю.И., Федотов А.М. Разработка единой информационной Системы СО РАН. //Материалы Международной научно-практической конференции "Развивающее образование XXI века, г. Горно-Алтайск, 2003 г., с. 258-262.

- [5] Федотов А.М. Технологии построения распределенных систем // Труды VIII Междунар. конф. по электронным публикациям. – Новосибирск. – 2003. (электронное издание, № гос. рег. 3521).
- [6] Шокин Ю.И., Барахнин В.Б., Пищик Б.Н., Ламин В.А., Жижимов О.Л., Федотов А.М., Мазов Н.А., Покровский Н.Н. Распределенная информационная система “Виртуальный музей науки и техники СО РАН” // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. V Всерос. науч. конф. RCDDL'2003. – Санкт-Петербург, НИИ Химии СПбГУ. – 2003. – С. 112-116. ил. ISBN 5-94158-056-8.
- [7] Шокин Ю.И., Федотов А.М. Создание и поддержка распределенных информационно-вычислительных ресурсов // В сб. Региональные проблемы перехода к устойчивому развитию. – Кемерово. – 2003.
- [8] Шокин Ю.И., Федотов А.М., Жижимов О.Л., Мазов И.А. Интегрированная распределенная информационная система (ИРИС) Сибирского отделения РАН // В сб.: Выездное заседание научно-координационного совета по целевой программе "Информационно-телекоммуникационные ресурсы СО РАН", Иркутск, 29-30 августа 2002, ИДСТУ СО РАН, 2003.
- [9] Жижимов О.Л. Введение в Z39.50: изд. 3-е доп. и перераб. // Новосибирск, Изд-во НГОНБ, 2003, 263 с., ил., табл.
- [10] Мазов Н.А., Жижимов О.Л. СИМІ (Z39.50) - основа интеграции информационных ресурсов по культурному наследию // Труды VIII Междунар. конф. по электронным публикациям. – Новосибирск. – 2003. (электронное издание, № гос. рег. 3521).
- [11] Бычков И.В, Гуськов А.Е., Жижимов О.Л., Мазов Н.А., Молородов Ю.И., Москвичев В.В., Федотов А.М. Распределенная информационная система "Виртуальный музей науки и техники СО РАН" // Материалы Международной конференции EVA-2003, 1-5 декабря 2003, г. Москва.
- [12] Паринев С.И., Ильменский М.Д., Маракуев А.В. Применение новых информационных технологий в экономических исследованиях. Журнал Экономика и математические методы, 2003, т. 39, номер 2, стр. 155-166.